SFAMNet: A Scene Flow Attention-based Micro-expression Network

Gen-Bing Liong[a], Sze-Teng Liong[b], Chee Seng Chan[a,*], John See[c]

[a] *CISiP, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia*
[b] *Department of Electronic Engineering, Feng Chia University, Taichung 40724, Taiwan R.O.C.*
[c] *School of Mathematical and Computer Sciences, Heriot-Watt University Malaysia, Putrajaya, Malaysia*

Abstract

Tremendous progress has been made in facial Micro-Expression (ME) spotting and recognition; however, most works have focused on either spotting or recognition tasks on the 2D videos. Until recently, the estimation of the 3D motion field (*a.k.a* scene flow) for the ME has only become possible after the release of the multi-modal ME dataset. In this paper, we propose the first Scene Flow Attention-based Micro-expression Network, namely SFAMNet. It takes the scene flow computed using the RGB-D flow algorithm as the input and predicts the spotting confidence score and emotion labels. Specifically, SFAMNet is an attention-based end-to-end multi-stream multi-task network devised to spot and recognize the ME. Besides that, we present a data augmentation strategy to alleviate the small sample size problem during network learning. Extensive experiments are performed on three tasks: (i) ME spotting; (ii) ME recognition; and (iii) ME analysis on the multi-modal CAS(ME)^3 dataset. Empirical results indicate that depth is vital in capturing the ME information and the effectiveness of the proposed approach. Our source code is publicly available at `https://github.com/genbing99/SFAMNet`.

Keywords: Facial micro-expression, scene flow, attention, spotting, recognition, analysis

1. Introduction

Facial expression is a form of non-verbal communication that conveys a person's emotion that words might not contain [1, 2]. Succinctly, there are two types of facial expressions, *i.e.*, Macro-Expression (MaE) and Micro-Expression (ME). It is known that analyzing ME is more challenging compared to the MaE in terms of the following characteristics: (i) *behavior* - ME is subtle and involuntary whereas MaE is noticeable and voluntary; (ii) *truthfulness* - ME occurs when a person tries to conceal genuine emotion whereas MaE states the person's actual feeling; (iii) *appearance* - ME may appear only on one region of the face whereas MaE appears across the entire face; and (iv) *timing* - ME usually last shorter than 0.5 seconds whereas MaE lasts between 0.5 to 4 seconds [3]. Within the occurrences of ME in a fraction of a second, a person that is professionally trained using the Micro-Expression Training Tool (METT) [4] can only correctly spot and recognize the ME around 50-50 chance [5], not to mention other people without professional training. It is noticed that an automatic ME analysis system has caught the eyes of both computer science and psychology experts. This is due to the increasing publicly available ME datasets and potential use cases in many sensitive fields, *i.e.*, police interrogation, clinical diagnosis, and law enforcement [6].

In the past decades, MEs have been studied vigorously in the 2D plane through image or video processing. However, 2D-based analysis is challenging to handle the subtle facial changes expressed on a 3D surface. Generally, it is often stated that 3D facial analysis provides more significant advantages in dealing with expression variations, pose variations, and viewpoint dependency [7]. Recently, [8] conducted a psychological experiment to validate that the third dimension (depth) is an essential factor by showing the subjects both 2D and 3D ME videos. The experiment found that the human visual perception system is more sensitive to 3D visualization with shorter reaction time and higher intensity ratings. Therefore, 3D motion-based ME analysis that takes additional depth information is presumed to reveal the visual clue that is not captured in a 2D video.

Typically, ME analysis comprises two main tasks: spotting and recognition. Influenced by the Micro-Expression Grand Challenge (MEGC) 2018–2022 [9, 10, 11, 12, 13], researchers tend to focus on only a single task, either spotting or recognition. Thus, the existing methods proposed for spotting task is difficult to be employed for recognition task, and vice versa. Albeit the Micro-Expression Analysis Network (MEAN) proposed by [14] can combat the problem, the two-step network learning paradigm is relatively time-consuming. Hence, developing an end-to-end learning approach that works well on both ME spotting and recognition tasks is essential.

A considerable amount of work has been established to

[*] Corresponding authors

Email addresses: `genbing67@gmail.com` (Gen-Bing Liong), `stliong@fcu.edu.tw` (Sze-Teng Liong), `cs.chan@um.edu.my` (Chee Seng Chan), `j.see@hw.ac.uk` (John See)

analyze the MaE in a 3D space [15]; however, minimal attempts have been proposed to analyze ME using 3D information. Until recently, such research direction was only made possible after the availability of multi-modal ME datasets [8, 16]. In this paper, we present SFAMNet, a Scene Flow Attention-based Micro-expression Network designed to spot and recognize the ME using color and depth information. SFAMNet takes the scene flow features as input and employs an attention-based end-to-end multi-stream multi-task architecture. We address the challenge of limited training data by proposing a data augmentation with an enhanced pseudo-labeling technique. Our main contributions are three-fold:

1. To the best of our knowledge, this is the first known work that attempts both spotting and recognition of MEs on the multi-modal CAS(ME)^3 dataset. We propose SFAMNet to accomplish both tasks as well as a unified spot-then-recognize analysis task.
2. Technically, we exploit the RGB-D flow feature extraction technique to estimate the 3D motion changes on the face using both color and depth modalities. Also, we propose a data augmentation strategy to increase the ME sample size for network training with an improved pseudo-labeling technique.
3. Experimentally, we achieve the best results in three tasks on the CAS(ME)^3 dataset: ME spotting with an F1-score of 0.0716, ME recognition (4-class) with UF1 of 0.4462, and ME analysis (4-class) with STRS of 0.0331.

2. Literature Review

Conventionally, the research on ME began in 1969 when Ekman discovered the occurrences of ME [17]. Then, the earliest posed ME datasets released in 2009 [18] spurs the development of computer vision algorithms. Since the posed MEs may present different characteristics (*i.e.*, behavior and timing) from the MEs in reality, the first spontaneous ME dataset was released in 2013 [19]. With additional depth information, the state-of-the-art multi-modal datasets were only released in the mid of 2022 to encourage the proposal of 3D-based techniques [8, 16]. Thereby, it is still considered very new with limited literature. Hence, this section reviews the methods proposed to deal with 2D videos.

2.1. Feature Extraction

The existing feature extraction techniques can be divided into two main categories: Local Binary Pattern (LBP) variants and optical flow guided features. LBP is one of the pioneer methods in the ME domain that is robust towards illumination changes and relatively simpler to be computed [20]. After that, the LBP on Three Orthogonal Plane (LBP-TOP) is commonly used as the baseline method in the ME datasets [19, 21, 22, 16] due to its ability to measure the intensity with an additional time dimension. Several LBP variants are then proposed to improve its discriminative power further [23, 24] However, the main drawback of most LBP variants is they only consider local instead of global features.

Meanwhile, the optical flow feature extraction technique is widely used in the ME domain [11] with its compelling ability to estimate the velocities of facial motion changes in a 2D field. There are a few optical flow computation algorithms including Horn & Schunck [25], Lucas Kanade [26], Farneback [27], and TV-L1 [28]. Then, the optical strain is derived from optical flow to capture the facial deformation information [29]. In the literature, some optical flow guided features are then designed to improve the robustness [30, 31, 32, 33]. It is also noticeable that there is a rising trend of using optical flow guided features in current literature.

2.2. ME Spotting

Generally, the ME spotting is performed after the feature extraction stage to locate the interval of the occurrences. In the work of [20], they proposed the feature difference (FD) analysis that computes the Chi-Square (χ^2) distance of the LBP features between two frames to obtain the level of facial changes. Afterward, thresholding and peak detection are applied to spot the relevant movements with a high change level [34, 35]. Furthermore, [36] analyzed the ME movement pattern and employed the Hammerstein model to augment the feature samples for training, leading to improved performance for the machine learning algorithm. A few traditional methods also analyze the optical flow intensity, which prevailed in the MEGC spotting task [37, 38, 39], but these methods are susceptible to large global movement [8].

Instead of directly using the features for the spotting task, deep learning-based methods are in fashion to further learn the salient regions contributing to the ME occurrences. Recurrent Neural Network (RNN) is used to learn the features sequentially [40]. Alternatively, several works formulate the learning task on frame level basis by designing innovative Convolutional Neural Network (CNN) architecture [41, 35, 42, 43, 44]. Interestingly, these CNN-based models take optical flow guided features as the input image, which again warrants the effectiveness of motion features.

2.3. ME Recognition

The ME recognition task is undoubtedly a classification problem. In the early works, the SVM classifier is frequently used to analyze the pattern from various features for comparison [30]. In tandem with the increasing ME samples in the literature, deep learning algorithms are garnering attention to learn the visual representation of emotions.

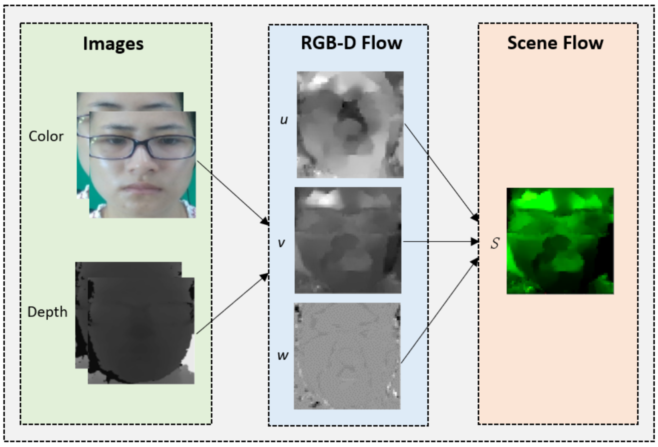


Figure 1: Implementation of RGB-D flow [51] to estimate the scene flow from two color and depth images.

The MEGC 2019 [10] has established a standard evaluation to measure the performance of the ME recognition task. In this challenge, the top 3 accepted submissions [45, 46, 47] utilized the optical flow features as the input data into the CNN architecture. Then, Feature Refinement (FeatRef) [48] was proposed to further learn the expression-specific motion features information. The Identity-aware and Capsule-Enhanced Generative Adversarial Network (ICE-GAN) [49] introduced the ME synthesis to augment the samples, meantime outperforming the existing works significantly. However, the loss functions used in the work are not disclosed entirely in the code repository for reproducibility.

2.4. ME Analysis

Despite numerous works proposed in the literature to perform ME spotting and recognition, the ME analysis task to spot-then-recognize the ME seamlessly remains subdued. This is mainly because the results from the spotting task are still far from satisfactory [13]. The prior research by [31] and [50] suggested a completely different process for spotting and recognition tasks, which requires additional efforts to be implemented. Thus, the Micro-Expression Analysis Network (MEAN) [14] has been devised to tackle both tasks in a single network. MEAN is a multi-output network with two task-specific networks for spotting and recognition. The imperfection of the network is the two-step learning paradigm, which can be time-consuming to train. To overcome this issue, an end-to-end learning approach is proposed in this paper.

3. Proposed Approach

3.1. Scene Flow

Scene flow estimates the 3D motion field between a pair of color (RGB) and depth (D) frames. It has been widely used in several applications, *e.g.*, autonomous driving, robotics, object segmentation, and many more [52]. Notwithstanding the ability to describe real-world motion, it has yet to be applied in the ME domain. In this paper, we implement the RGB-D flow [51] to reveal the clues of 3D facial motion using two RGB-D images, as shown in Figure 1.

Given I_0, I_1 be the intensity images and Z_0, Z_1 be the depth images, with similar height H and width W, taken at time t_0, t_1, respectively. According to the brightness constancy implied by the optical flow, where the moving points do not move between the time t_0 and t_1, we can formulate the brightness constancy term:

$$E_C = I_1(x+u, y+v) - I_0(x, y) \tag{1}$$

where x, y are the pixel coordinates and u, v are the horizontal and vertical components of the flow field. Formally, it is assumed that the depth of the moving points is not constant over time, but the depth change must be equal to the difference between Z_0 and Z_1 that warped with the optical flow. Hence, the depth constancy term can be defined as:

$$E_Z = Z_1(x+u, y+v) - Z_0(x, y) - w \tag{2}$$

where w refers to the depth component of the flow.

In practice, the aperture problem is associated with scene flow estimation. To solve this, regularization is required to provide a smooth flow field. The regularization term based on total variation can be formulated as:

$$E_R = |\nabla u|^2 + |\nabla v|^2 + \beta(x, y)|\nabla w|^2 \tag{3}$$

where $\beta(x, y) = f^2$, and f is the focal length of the color camera. Furthermore, the flow magnitude penalty term is defined as:

$$E_P = |{}^{f}/{}_{Z(x,y)} \cdot (u, v, w)|^2 \tag{4}$$

To complete the scene flow estimation, the objective is to minimize the energy function:

$$E = E_C + E_Z + \alpha E_R + \gamma E_P \tag{5}$$

where α and γ are the constant weights that can be tuned.

Finally, Euler-Lagrange equations with Successive Over-Relaxation (SOR) updates come into place to solve the energy minimization problem, following the optimization process proposed in the work of [51]. As a result, the scene flow can be represented as $\mathcal{S} = (u, v, w) \in \mathbb{R}^{W \times H \times 3}$. It is worth mentioning that the scene flow consumes significantly lesser memory than the depth flow $\mathcal{D} \in \mathbb{R}^{W \times H \times 200}$ introduced in the work of [8].

3.2. Data Augmentation

Insufficient training data is a long-standing issue in the ME domain [53, 8]. Hence, we propose a data augmentation strategy that utilizes the sliding window approach to generate new training data from existing data. As illustrated in Figure 2, the training data and labels computed using ground-truth and sliding window approaches are combined to facilitate the network training process.

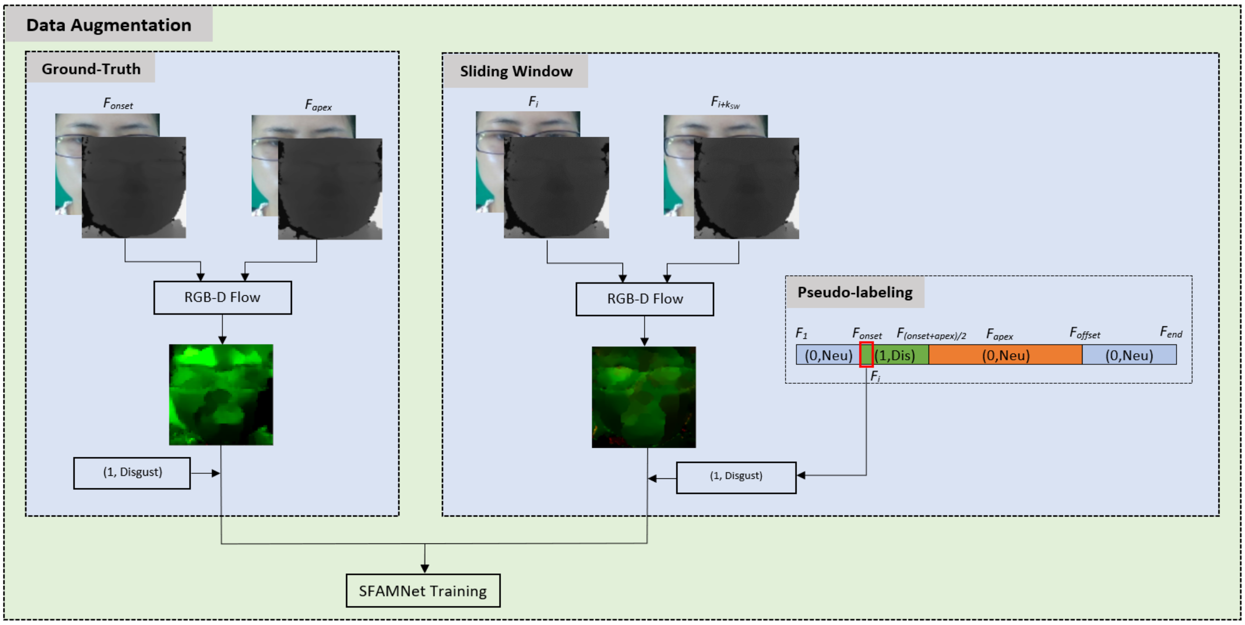


Figure 2: Data augmentation strategy that concatenates the training data generated from ground-truth and sliding window approaches.

Given a multi-modal ME dataset containing RGB and D information for each frame recorded. Generally, the annotation of frame location (*e.g.*, F_{onset}, F_{apex}, and F_{offset}) and emotion class (*e.g.*, happy, disgust, ..., others) labels are provided for each ME clip. For the ground-truth data, RGB-D flow is employed to compute the scene flow for each pair (F_{onset}, F_{apex}). Then, we can generate the ground-truth label set, such that:

$$Y_{GT} = (s_i, c_i) \text{ for } i = 1, \ldots, n \tag{6}$$

where s_i is the spotting confidence score and c_i is the emotion class for the i clip, while n is the total number of ME clips in the dataset. Note that s_i is always set to 1 for the ground-truth labels.

For the sliding window approach, the window with the interval $[F_i, F_{i+k_{SW}}]$ is scanned across each video to compute the scene flow, where k_{SW} is the frame distance that is computed based on the video FPS, such that: $k_{SW} = FPS \times 0.2$. The 0.2s is the upper limit duration of the onset phase (onset until apex) [54]. To label the scene flow for training, we revisit and enhance the pseudo-labeling technique proposed in the work of [35]. Concretely, we initialize the pseudo-label set with a similar length to the video as ($s_i = 0, c_i = Neutral$). Intuitively, we find that the AUs produced in the interval $[F_{onset}, F_{(onset+apex)/2}]$, which captures the initial activation of the expression, tend to provide the most valuable insights for assessment. Hence, we label the frames in the interval of each ME clip as ($s_i = 1, c_i$). To ensure that the pseudo-labeling process remains within the bounds of the video length, we only consider the first frame of the window for labeling. Thus, the pseudo-label set can be formulated as:

$$Y_{SW} = (s_i, c_i) \text{ for } i = 1, \ldots, F_{end} \tag{7}$$

where F_{end} is the last frame of the video, and the process is repeated for all videos in the dataset.

Finally, the training labels for the network input can be obtained by concatenating the ground-truth label set and pseudo-label set:

$$Y = Concat(Y_{GT}, Y_{SW}) \tag{8}$$

Since some existing works [35, 55] only apply the sliding window approach for the ME spotting task, theoretically, our proposed approach which incorporates the actual ME samples from ground-truth labels can describe the motion details more accurately. Likewise, most works only use ground-truth labels to compute features for the ME recognition task. As a side benefit, our implementation of the sliding window approach increases the ME samples in the same direction as the ground-truth motion field with different magnitudes.

3.3. SFAMNet

Our proposed Scene Flow Attention-based Micro-expression Network, namely SFAMNet, is an attention-based end-to-end multi-stream multi-task network, as depicted in Figure 3. Specifically, the three inputs of the network are the scene flow components (u, v, w), and the two outputs are the confidence score for spotting and the emotion class for recognition. Although the architecture of SFAMNet is based on MEAN [14], several significant refinements distinguish it from its predecessor:

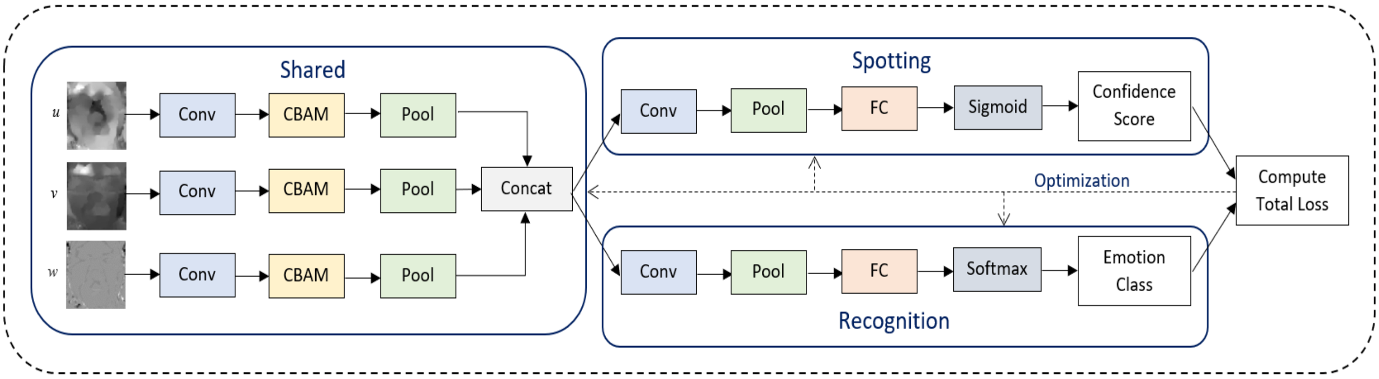


Figure 3: Our proposed SFAMNet takes the scene flow components (u, v, w) into each stream as inputs, and predicts the confidence score (spotting) and emotion class (recognition).

1. The inputs to the network streams are the scene flow components (u, v, w), which describe the motion information with an extra depth dimension.
2. The Convolutional Block Attention Module (CBAM) structure seamlessly integrates into the network architecture to focus more on the salient regions that contribute to the ME.
3. To eliminate the cumbersome two-step training paradigm, an end-to-end network optimization method is proposed to simplify the training process.
4. The number of filters for the three-stream convolutional layers is standardized to three.
5. At the spotting network, the linear activation function is replaced with the sigmoid activation function.

Given the three scene flow components (u_i, v_i, w_i) of i-th frame fed into each stream of the inputs, our proposed SFAMNet model $\mathcal{M}$ predicts the confidence score $(\hat{s}_i)$ and emotion class $(\hat{c}_i)$, which can be represented as:

$$\hat{s}_i, \hat{c}_i = \mathcal{M}(u_i, v_i, w_i) \text{ for } i = 1, \ldots, F_{end} \tag{9}$$

where F_{end} is the last frame in the video.

There are three main modules in the SFAMNet architecture: (1) shared network, which is integrated with CBAM to extract the features from each scene flow component; (2) spotting network, which outputs the confidence score for spotting task; (3) recognition network, which predicts the emotion class for the recognition task.

3.3.1. CBAM

The Convolutional Block Attention Module (CBAM) [56] is in fashion due to its capability to emphasize meaningful features using the attention mechanism. More importantly, CBAM can be integrated into any CNN architecture and is end-to-end trainable. There are two modules placed sequentially, *i.e.*, channel attention module and spatial attention module. The details are discussed next.

The channel attention module concentrates on "what" is important in the image by computing the channel-wise attention. First, the global Max Pooling (MaxPool) and global Average Pooling (AvgPool) are applied simultaneously to squeeze the input feature map F. Specifically, AvgPool aggregates the spatial information, whereas MaxPool preserves the contextual information in the feature map. Next, the feature maps F^c_{avg} and F^c_{max} obtained are forwarded to a shared Multi-Layer Perceptron (MLP) with one hidden layer, the weights in the layer are denoted as W_0 and W_1. Afterwards, the features are summed up element-wise and passed through a sigmoid function (σ) to get the output channel attention map M_c. Overall, the channel attention module can be summarized as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \end{aligned} \tag{10}$$

To broadcast the attention values computed, the $M_c(F)$ is then multiplied element-wise with input features F to form F':

$$F' = M_c(F) \otimes F \tag{11}$$

On the other hand, the spatial attention module focuses on "where" is the important region, as complementary to the channel attention. Taking the intermediate feature maps F' from the output of the channel attention module, the global AvgPool and global MaxPool are applied to obtain the 2D feature maps: F^s_{avg} and F^s_{max}. After concatenating the 2D feature maps, a convolutional layer with 7 × 7 kernel size $(f^{7\times7})$ and a sigmoid function (σ) are then applied to generate a spatial attention map M_s. The process can be written as:

$$\begin{aligned} M_s(F') &= \sigma(f^{7\times7}([AvgPool(F'); MaxPool(F')])) \\ &= \sigma(f^{7\times7}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{12}$$

Finally, the $M_s(F')$ is multiplied element-wise with F' to get the refined output:

$$F'' = M_s(F') \otimes F' \tag{13}$$

Formally, given the feature map $F \in \mathbb{R}^{C_m \times H_m \times W_m}$ as input, CBAM generates the $M_c \in \mathbb{R}^{C_m \times 1 \times 1}$ and $M_s \in \mathbb{R}^{1 \times H_m \times W_m}$ as attention maps, then $F'' \in \mathbb{R}^{C_m \times H_m \times W_m}$ as the final output. Note that C_m, H_m, and W_m denote the channel, height, and width of the feature map, respectively.

3.3.2. Shared Network

A shared network comprises three input streams, each taking a scene flow component. First, the convolutional layer applies convolutional filters to compute a feature map, such that every pixel z_{xy} in the feature map at position (x, y) of the l-th layer is computed by:

$$f_{conv}(z^l_{xy}) = \sum_{p=0}^{P^l-1} \sum_{q=0}^{Q^l-1} w^l_{pq} z^{(l-1)}_{(x+p)(y+q)} + b^l \tag{14}$$

where w^l_{pq} is the kernel weight parameter at position (p, q) of layer l, P^l and Q^l are the width and height of the kernel in layer l respectively, b^l is the bias in layer l. Subsequently, the CBAM layer is integrated to extract meaningful features in both channel and spatial dimensions. Since the size of input and output feature maps are the same, for simplicity, we denote it as follows:

$$f_{cbam}(z) = cbam(z) \tag{15}$$

Next, we use the max pooling layer to calculate the maximum value in each patch of the feature map, which can be expressed as:

$$f_{pool}(z^l_{xy}) = max\{z^{(l-1)}_{(x+j)(y+k)} | j=0, \ldots, m; k=0, \ldots, m\} \tag{16}$$

where m refers to the patch size. Lastly, the outputs from three streams are merged using the concatenation layer (f_{concat}) to represent the features extracted from the input scene flow. Succinctly, the shared network can be represented as:

$$\begin{aligned} F(\tau) &= f_{pool}(f_{cbam}(f_{conv}(\tau))) \\ F_s(u, v, w) &= f_{concat}(F(\tau)) \text{ for } \tau = u, v, w \end{aligned} \tag{17}$$

where $F(\tau)$ is the intermediate feature map before the concatenation layer and F_s is the output of the shared network.

3.3.3. Spotting Network

A spotting network is a task-specific network designed for the spotting task. After the shared network, a convolutional layer followed by a max pooling layer is applied to the feature maps to extract the important spotting task-specific features. Afterwards, the feature maps are flattened into a one-dimensional vector, then passed on to the fully connected layer (f_{fc}) with a sigmoid activation function (f_{sig}) that has only one neuron. The sigmoid activation function can be computed as:

$$f_{sig}(z) = \frac{1}{1 + e^{-z}} \tag{18}$$

The output is the predicted spotting confidence score ($\hat{s}$) between 0 and 1, which can be interpreted as the probability of being in an interval of expression. The spotting network is as follows:

$$\hat{s}(F_s) = f_{sig}(f_{fc}(f_{pool}(f_{conv}(F_s)))) \tag{19}$$

where $\hat{s}(F_s)$ is the predicted spotting confidence score from the output of the shared network.

3.3.4. Recognition Network

The recognition network takes the feature maps after the shared network as input and predicts the probabilities for emotion classes. It has a similar structure as the spotting network until the fully connected layer. In contrast, the features extracted are specific to the recognition task. A softmax activation function (f_{soft}) is appended, such that:

$$f_{soft}(z_j) = \frac{e^{z_j}}{\sum_{c=1}^{C} e^{z_c}} \text{ for } j = 1, \ldots, C \tag{20}$$

where z_j is the probability of the feature map z to be class j, and C is the total number of emotion classes. The output of the recognition network is the predicted emotion class ($\hat{c}$). The network is structured as follows:

$$\hat{c}(F_s) = f_{soft}(f_{fc}(f_{pool}(f_{conv}(F_s)))) \tag{21}$$

where $\hat{c}(F_s)$ is the predicted probability for each emotion class from the output of the shared network.

3.3.5. Network Optimization

Our proposed network is trained end-to-end, which requires an optimization process to handle the predicted multi-task outputs $(\hat{s}_i, \hat{c}_i)$. The training labels (s_i, c_i) generated from Equation 8 are used herein as the actual labels to optimize the network.

In particular, the spotting loss $\mathcal{L}_{spot}$ is designed to reduce the Mean Squared Error (MSE) between the predicted and actual spotting confidence score:

$$\mathcal{L}_{spot} = \frac{1}{n} \sum_{i=1}^{n} (s_i - \hat{s}_i)^2 \tag{22}$$

where n is the total number of training samples.

Since the recognition network predicts a multi-class output ($\hat{c}_i$), the actual emotion label (c_i) is converted to a one-hot vector. Hence, we can formulate the recognition loss $\mathcal{L}_{recog}$ using the categorical cross-entropy loss to measure how well the predicted probability for each emotion class:

$$\mathcal{L}_{recog} = -\frac{1}{n} \sum_{i=1}^{n} c_i \, log\,(\hat{c}_i) \tag{23}$$

The final objective is to minimize the total loss:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{spot} + (1 - \lambda) \mathcal{L}_{recog} \tag{24}$$

where λ is a balancing parameter between 0 and 1.

3.4. Prediction and Post-processing

During the prediction of a long video, each frame with extracted scene flow features is fed into the network to obtain the predicted spotting confidence score and probability of emotion labels. Then, we select the emotion label with the largest probability as the predicted emotion class.

During post-processing for the spotting task, we implement the strategy proposed in the work of [57]. Notably, this strategy can obtain both MaE and ME intervals for evaluation. In short, there are five main steps: (i) get the predicted confidence score for each frame in the entire video; (ii) use two sliding windows with length k_{MaE} and k_{ME} and apply a simple average smoothing function to obtain two smoothed curves; (iii) peak detection technique with a threshold of value p is utilized to detect the peak frames in both the ME and MaE curves; (iv) difference between the spotted ME peak frame, and its immediate frame in the MaE curve is computed, then, the ME peak frame is kept if the difference is larger than the threshold d. This process aims to reduce a large amount of falsely spotted ME from the previous step; and (v) the intervals of MaE and ME are spotted based on the length of α_{MaE}^{onset}, α_{MaE}^{offset}, α_{ME}^{onset}, and α_{ME}^{offset}.

During post-processing for the analysis task, we adopted the mode technique [14] to determine the emotion class of the spotted interval. Briefly, the highest number of occurrences of the predicted emotion class in the spotted onset phase is selected. Accordingly, this mode technique reduces the bias (using only a single frame) and noise (using too many frames) from the entire spotted interval.

4. Experiment

4.1. Dataset

We select the multi-modal CAS(ME)^3 dataset [8] to evaluate the performance of our proposed approach. CAS(ME)^3 is the first ME dataset that introduces the depth information to describe the facial depth changes during an expression. The Intel RealSense® D415™ camera with 30 FPS and 1280 × 720 resolution is used to record the RGB-D images. The dataset has three parts: part A, part B, and part C. The male:female ratio is 112:135, with a mean age of 22.74. Part A and part B are recorded using the second-generation elicitation paradigm, while part C is based on the third-generation ME elicitation paradigm. Due to the videos in part B being unlabeled, and videos in part C being elicited in a different setting; hence, we use part A for experiments.

Part A is in the second generation, which is recorded in a constrained lab environment. In order to capture the spontaneous ME, the subjects are asked to keep a poker face throughout the process. There are 1300 video clips collected from 100 subjects, and the average video duration is around 98 seconds. A total of 860 MEs and 3342 MaEs are annotated with onset, apex, and offset frames location. The emotion and AU labels are provided only for MEs. The emotion can be classified into 4 classes (457 negative, 55 positive, 187 surprise, and 161 others) or 7 classes (250 disgust, 187 surprise, 161 others, 86 fear, 64 anger, 57 sad, and 55 happy).

4.2. Performance Metrics

To assess the performance of our proposed approach, we follow the metrics suggested in the work of [14] for evaluating long videos on spotting, recognition, and analysis tasks.

For the spotting task, we adopt the F1-score metric as suggested by the ME community [11, 12, 13]. We employ the standard evaluation metrics for recognition and analysis tasks, such as recall, precision, F1-score, UF1, and UAR [10]. Note that the recognition task is based on ground-truth intervals, whereas the analysis task depends on spotted TP intervals. We use the Spot-Then-Recognize Score (STRS) to measure the overall system performance, which multiplies the F1-score from the spotting and analysis tasks [14].

4.3. Experiment Settings

Before computing the scene flow, we crop the facial region using Multi-task Cascaded Convolutional Networks (MTCNN) [58]. Besides, the global motion removal [35] is applied by deducting the scene flow of the nose region. The three flow components are then resized to 42 × 42 pixels before being fed into the proposed network. For intelligent analysis during the spotting task, we have excluded some MEs that have long offset phases (apex until offset), as suggested by the authors of the target dataset [8].

Our network is trained with Adam optimizer and the batch size is set to 1024. We employ a learning rate of 1 × 10^{-4} and execute for 200 epochs. To alleviate the class imbalance problem, we sample the expression and non-expression frames with a ratio of 1:1. Also, we balance the class weights according to the number of emotion class samples to regularize the recognition loss in Equation 23. By applying more weight to the minority class and less weight to the majority class, the network is forced to learn the representations equally for all classes. To ensure the predicted emotion class is valid, we set the class weight of class 'neutral' to 0. We first train the network on 4-class evaluation. Subsequently, we freeze the shared network and spotting network and fine-tune it on 7-class evaluation.

The Leave-One-Subject-Out (LOSO) cross-validation is used to evaluate the proposed approach. This eliminates the subject bias by ensuring the model has no knowledge about the testing subject. For parameter settings, we set

Table 1: Performance comparison for ME spotting on CAS(ME)^3 dataset. Results #1∼#3 and #6 are obtained from [8]. $\mathcal{O}$: Optical flow, $\mathcal{D}$: Depth flow, $\mathcal{S}$: Scene flow, $\mathcal{H}$: Hybrid flow

#	Authors	Features	Method	F1-score
1	He [37]	$\mathcal{O}$	OF-FD	0.0000
2	Zhang et al. [59]	$\mathcal{O}$	SP-FD	0.0103
3	Yu et al. [42]	$\mathcal{O}$	LSSNet	0.0653
4	Liong et al. [14]	$\mathcal{O}$	MEAN	0.0283
5	Ours	$\mathcal{O}$	SFAMNet	0.0591
6	Li et al. [8]	$\mathcal{D}$	SP-FD	0.0112
7	Liong et al. [14]	$\mathcal{S}$	MEAN	0.0412
8	Ours	$\mathcal{S}$	SFAMNet	0.0695
9	Ours	$\mathcal{H}$	SFAMNet	**0.0716**

the k_{SW}, k_{ME}, k_{MaE}, α_{ME}^{onset}, α_{ME}^{offset}, α_{MaE}^{onset}, α_{MaE}^{offset}, p, d to 6, 15, 20, 15, 32, 50, 50, 0.58, and 0.02, respectively. The balancing parameter λ during network optimization is set to be 0.9.

All experiments are performed with Pytorch on NVIDIA GeForce RTX 3090. Our proposed SFAMNet is a shallow architecture with 7237 parameters and 2.8 million FLOPs. The average time taken for a single fold of LOSO on the CAS(ME)^3 dataset is around 52.83s.

5. Results And Discussion

This section comprehensively discusses the results for (i) ME spotting, (ii) ME recognition, and (iii) ME analysis. Besides our proposed SFAMNet with scene flow features, we also conducted two additional experiments: SFAMNet with optical flow features and SFAMNet with *hybrid* flow features. The hybrid flow features are obtained by combining the u and v components from optical flow with the w component from scene flow. The intuition behind the hybrid flow features is to leverage the complementary information provided by both optical flow and scene flow; this is explained in Section 5.4 (flow components) in more detail.

5.1. ME Spotting

The results for ME spotting are reported in Table 1. Our proposed approach with the hybrid flow features (#9) outperforms other methods (#1∼#8) by achieving the F1-score of 0.0716. Comparing the approaches between optical flow features (#5) and scene flow features (#8), it is observed that scene flow performs better because of the incorporation of additional depth modality to capture the facial micro-movement in 3D space. Therefore, we prove that the expansion from 2D to 3D in the motion field is important to describe the subtle movement of ME in all directions (*i.e.,* horizontal, vertical, and depth).

It is worth mentioning that the OF-FD [37] (#1) is the top-1 method from the MEGC 2021 spotting task, however, it is discerned that no TP is spotted. Besides, the top-1 method from MEGC 2020 spotting task, which is SP-FD [59] (#2) has been reported. Instead of using the optical flow features, the authors of CAS(ME)^2 dataset [8] (#6) re-implemented the SP-FD method with the depth flow features. Nevertheless, both results were found to be unsatisfactory. As a whole, the traditional FD methods have weaker generalization ability and require explicit parameter settings. On the contrary, the deep learning approaches (#3∼#5 and #7∼#9) performed significantly well. This could be ascribed to the large sample size that facilitates the network to learn the characteristics of the ME occurrences better. While the top-2 winner of MEGC 2021, which is LSSNet [42] (#3) performs slightly better than our SFAMNet (#5) on the ME spotting task when using optical flow features as input, SFAMNet still demonstrates strong performance and has an additional capability to perform multi-task, *i.e.*, spotting and recognition. Interestingly, it is found that our SFAMNet (#5 and #8) outperforms MEAN [14] (#4 and #7) quite significantly either with the optical flow or scene flow features as input. This highlights the robustness of our method after the network architecture refinements.

5.2. ME Recognition

We compare the performance of the benchmark results, our proposed approach, and our additional experiments for the ME recognition task, as shown in Table 2. Our proposed approach with the hybrid flow features (#11) surpasses all other methods (#1∼#10) by obtaining the UF1 of 0.4462 and UAR of 0.4797 on 4-class evaluation, besides, showing considerable improvement compared to the baseline, re-implemented, and proposed methods with different features (#4∼#10) by achieving UF1 of 0.2365 and UAR of 0.2373 on 7-class evaluation. In particular, the experiment results suggest that using the motion information rather than the RGB-D image as network input improves the recognition performance. To emphasize the importance of scene flow features in our approach (#10), we conduct an experiment using RCN-A [60] (#8), which is the method that achieves the best recognition performance with RGB-D image as input. As a result, it is evident that scene flow can better represent the ME information as compared to optical flow by enabling the network to distinguish the emotions with additional facial depth information.

To understand each emotion class's difficulty from our proposed approach's (#10) prediction, we provide the confusion matrices of 4-class and 7-class in Figure 5. According to the 4-class analysis, the performance is slightly higher on negative and surprise emotions, with 0.5120 and 0.4599, respectively. This is attributed to the larger sample size of negative (457) and surprise (187) compared to the positive (55) and others (160). An identical scenario turns up on the 7-class analysis, the majority class such as disgust (250), surprise (187), others (160), fear (86), and anger (64) prevailed the minority class such as sad (57) and happy (55). Hence, gathering more ME samples

Table 2: Performance comparison for ME recognition of 4-class and 7-class on CAS(ME)^3 dataset. Results #1∼#4 are obtained from [8]. $\mathcal{O}$: Optical flow, $\mathcal{S}$: Scene flow, $\mathcal{H}$: Hybrid flow

#	Authors	Input/Features	Method	4-class		7-class	
				UF1	UAR	UF1	UAR
1	Liong et al. [46]	RGB-D Image	STSTNet	0.3795	0.3792	-	-
2	Xia et al. [60]	RGB-D Image	RCN-A	0.3928	0.3893	-	-
3	Zhou et al. [48]	RGB-D Image	FeatRef	0.3493	0.3413	-	-
4	Li et al. [8]	RGB-D Image	AlexNet	0.3001	0.2982	0.1773	0.1829
5	Xia et al. [60]	$\mathcal{O}$	RCN-A	0.4002	0.4229	0.1892	0.1879
6	Liong et al. [14]	$\mathcal{O}$	MEAN	0.3894	0.4004	0.1994	0.1929
7	Ours	$\mathcal{O}$	SFAMNet	0.3853	0.4142	0.1778	0.1860
8	Xia et al. [60]	$\mathcal{S}$	RCN-A	0.3966	0.3978	0.1994	0.2006
9	Liong et al. [14]	$\mathcal{S}$	MEAN	0.3968	0.4187	0.1916	0.2024
10	Ours	$\mathcal{S}$	SFAMNet	0.4006	0.4271	0.2000	0.2091
11	Ours	$\mathcal{H}$	SFAMNet	**0.4462**	**0.4797**	**0.2365**	**0.2373**

Table 3: Performance comparison for ME analysis of 4-class and 7-class on CAS(ME)^3 dataset. $\mathcal{O}$: Optical flow, $\mathcal{S}$: Scene flow, $\mathcal{H}$: Hybrid flow

#	Authors	Features	Method	4-class		7-class	
				F1-score	STRS	F1-score	STRS
1	Liong et al. [14]	$\mathcal{O}$	MEAN	0.3532	0.0100	0.1915	0.0054
2	Ours	$\mathcal{O}$	SFAMNet	0.4156	0.0149	0.1976	0.0117
3	Liong et al. [14]	$\mathcal{S}$	MEAN	0.4150	0.0171	0.2300	0.0095
4	Ours	$\mathcal{S}$	SFAMNet	0.4156	0.0289	**0.2358**	0.0164
5	Ours	$\mathcal{H}$	SFAMNet	**0.4619**	**0.0331**	0.2341	**0.0168**

for network training is necessary to boost the generalization ability and reduce the prediction bias based on the imbalanced dataset.

5.3. ME Analysis

In reality, the human perception system performs the ME spotting and ME recognition sequentially, where the occurrence is first spotted and then emotion is classified. However, most existing works struggle with the ME spotting task in long videos, leading to a lack of interest in developing a complete ME analysis system. In Table 3, we report the performance of our proposed approach against the only existing work that performed ME analysis on the long videos, which is MEAN [14]. Note that, we take only the spotted TP intervals for analysis. Our proposed approach with hybrid flow features (#5) outperforms the rest (#1∼#4) in terms of STRS, achieving an STRS of 0.0331 in the 4-class evaluation and an STRS of 0.0168 in the 7-class evaluation. Moreoever, it is noted that our SFAMNet (#2 and #4) outperforms MEAN (#1 and #3) in all metrics when using either optical flow or scene flow features as input. This verifies that our proposed network architecture is preferable to the MEAN architecture on the ME analysis task.

To investigate the performance improvement of scene flow over optical flow, it is observed that our proposed approach with scene flow (#4) performs either equally or better than the one with optical flow (#2) on both 4-class and 7-class evaluations in terms of F1-score and STRS. For a closer inspection, the confusion matrices of our approach with scene flow (#4) on 4-class and 7-class ME analysis are illustrated in Figure 6. Corresponding with the ME recognition results in Figure 5, the performance of positive and happy in 4-class and 7-class, respectively, are among the lowest compared to the rest. From the spotting task, it is found that there is only 1 happy (or positive) emotion out of 81 spotted ME samples. Therefore, we discern that the number of spotted ME samples for a particular emotion class correlates with its classification result.

Table 4: Performance comparison for MaE spotting on CAS(ME)^3 dataset. $\mathcal{O}$: Optical flow, $\mathcal{S}$: Scene flow, $\mathcal{H}$: Hybrid flow

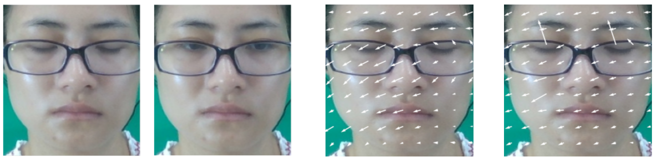
#	Features	Method	F1-score
1	$\mathcal{O}$	SFAMNet	0.1254
2	$\mathcal{S}$	SFAMNet	0.0854
3	$\mathcal{H}$	SFAMNet	**0.1484**

5.4. Ablation Studies

MaE Spotting: Since there was no previous attempt at MaE spotting on CAS(ME)^3 dataset, we report the results of our proposed approach and the additional experiments

Table 5: Performance comparison between different training data generation methods on CAS(ME)^3 dataset. G/T: Ground-truth; SW: Sliding Window

Method		G/T	SW	G/T+SW
# of training samples		1700	4472	**6172**
ME Spotting (F1-score)		0.0041	0.0420	**0.0695**
ME Recognition (UF1)	4-class	0.3403	0.3338	**0.4006**
	7-class	**0.2239**	0.1996	0.2000
ME Analysis (STRS)	4-class	0.0000	0.0112	**0.0289**
	7-class	0.0014	0.0106	**0.0164**



(a) Eye blinking action (b) Scene flow (c) Optical flow

Figure 4: Visualization of the motion field from both scene flow and optical flow during the eye blinking action.

in Table 4. Our proposed approach with hybrid flow features (#3) outperforms the rest (#1∼#2) by obtaining an F1-score of 0.1484. Besides, it is noticed that the optical flow features (#1) obtained a significantly higher F1-score than the scene flow features (#2), with a difference of nearly 47%. Upon careful inspection, the RGB-D flow method that computes the scene flow features is weak in handling large movements and missing objects [51]. As illustrated in Figure 4, the scene flow is unable to handle huge movement with missing pixels during the eye blinking action, in contrast to the optical flow computed using the TV-L1 algorithm [28].

Data augmentation: Experiments are conducted to determine the impact of data augmentation strategy on the network performance, as presented in Table 5. With different training data generation methods, it is observed that the combination of ground-truth and sliding window outperforms the others on all tasks except the 7-class ME recognition. This is attributed to the highest number of training samples which improves the network generalizability and reduces the chances of overfitting. Notwithstanding the number of training samples of ground-truth is lesser than the sliding window, the recognition result is slightly better due to the more precise ME location used to compute the feature maps. However, the spotting performance of ground-truth is worse than the sliding window because the sliding window approach generates a larger ME sample size with diverse ME patterns. This study ascertains our presumption that utilizing both ground-truth and sliding window approaches is complementary to each other.

Attention mechanism: To dive further into the attention mechanism, we apply the Gradient-weighted Class activation mapping (GradCam) to visualize the activated regions of SFAMNet with and without the integration of CBAM. As depicted in Figure 7, the coarse localization heatmaps are generated for the layers before and after the concatenation layer. Besides, the heatmaps are overlaid on both the feature map and raw image to visualize the activated regions. The colors range from red to blue, in which red indicates high activation and blue indicates low activation.

Taking the three scene flow components as input, the activation regions for each stream of the SFAMNet with CBAM before the concatenation layer are as follows: u — eyes and eyebrows; v — eyebrows; w — upper lid. After concatenation, the network emphasizes the eyebrows regions, which match with the ground-truth label AU 4 (brow lowerer). Hence, the network accurately predicts the emotion being 'Negative' and obtained UF1 of 0.4006. On the contrary, the SFAMNet without CBAM are as follows: u — eyes, eyebrows, and nose; v — eyebrows and eyes; w — upper lid. After concatenation, the network produces a larger activated region covering the entire upper face (including eyes, eyebrows, and nose regions). The predicted 'Others' emotion is incorrect and a lower UF1 of 0.3636 is obtained. To this end, we demonstrate the attention mechanism's significance and provide a visual explanation that our proposed network can localize the related AU.

Flow components: According to the results from Table 4, we find that the optical flow features as network inputs improve the performance of MaE spotting. Therefore, we conduct experiments by varying the network inputs with the components from optical flow and scene flow. Note that the optical flow is calculated using TV-L1 algorithm [28], whereas the scene flow is computed using RGB-D Flow [51]. As shown in Table 6, it is observed that the hybrid flow features with a combination of u, v, w, where u and v are from optical flow and w is from scene flow, consistently outperforms other combinations on all tasks. This indicates that the optical flow is robust to reveal the horizontal (u) and vertical (v) motion details, while scene flow is important in capturing the depth (w) motion details to improve the network performance further. Having said that, the computation of hybrid flow features requires the optical flow and scene flow features to be calculated separately, which is expensive.

On a side note, we compare the two inputs (u and v) from optical flow and scene flow. It is noticed that optical flow achieves a better result than scene flow on all tasks. This further confirms that optical flow is preferable for the computation of u and v components. In contrast, the performance of a single input (ϵ — derivative of optical flow) is slightly lower because less information is supplied for network training. Besides, we also see that taking the depth component (w) from scene flow as input achieved the lowest results, representing that using the depth information alone does not guarantee a convincing performance. Hence, it deserves further exploration to understand the

Table 6: Performance comparison of different flow components as inputs of the SFAMNet on CAS(ME)^3 dataset. $\mathcal{O}$: Optical flow; $\mathcal{S}$: Scene flow, $\mathcal{H}$: Hybrid flow

Flow components		$\mathcal{O}$			$\mathcal{S}$			$\mathcal{H}$
		$\mathcal{O}_u, \mathcal{O}_v$	$\mathcal{O}_\epsilon$	$\mathcal{O}_u, \mathcal{O}_v, \mathcal{O}_\epsilon$	$\mathcal{S}_u, \mathcal{S}_v$	$\mathcal{S}_w$	$\mathcal{S}_u, \mathcal{S}_v, \mathcal{S}_w$	$\mathcal{O}_u, \mathcal{O}_v, \mathcal{S}_w$
Spotting (F1-score)	ME	0.0490	0.0408	0.0591	0.0413	0.0000	0.0695	**0.0716**
	MaE	0.1441	0.1235	0.1254	0.0798	0.0018	0.0854	**0.1484**
	Overall	0.1195	0.1025	0.1079	0.0729	0.0014	0.0809	**0.1276**
ME Recognition (UF1)	4-class	0.4325	0.3390	0.3853	0.3468	0.2740	0.4006	**0.4462**
	7-class	0.2320	0.1731	0.1778	0.2045	0.1108	0.2000	**0.2365**
ME Analysis (STRS)	4-class	0.0190	0.0159	0.0149	0.0166	0.0000	0.0289	**0.0331**
	7-class	0.0099	0.1763	0.0117	0.0038	0.0000	0.0164	**0.0168**

Table 7: Performance comparison of different pseudo-labeling techniques for the spotting confidence score with scene flow as input.

Pseudo-labels		Hard (Ours)	Soft
Spotting (F1-score)	ME	**0.0695**	0.0440
	MaE	**0.0854**	0.0734
	Overall	**0.0809**	0.0669
ME Recognition (UF1)	4-class	**0.4006**	0.3691
	7-class	**0.2000**	0.1984
ME Analysis (STRS)	4-class	**0.0289**	0.0156
	7-class	**0.0164**	0.0078

importance of depth dimension when an expression occurs.

Pseudo-labeling: To gain insights into the effect of different spotting confidence scores, we have conducted an experiment to compare the performance of two pseudo-labeling approaches. The first one is the proposed method, where all confidence scores are labeled as 1 (referred to as hard pseudo-labels). The second method involves labeling the confidence score based on the distance from the apex (referred to as soft pseudo-labels). Specifically, for the frames located from the onset until the apex, they are labeled as: $s_i = F_i/(F_{apex} - F_{onset})$. Whereas for the frames located from the apex until the offset, they are labeled as: $s_i = (F_{offset} - F_i)/(F_{offset} - F_{apex})$. The results of these two pseudo-labeling methods are compared in Table 7. It is observed that the hard pseudo-labels outperform the soft pseudo-labels quite significantly across all tasks. This suggests that assigning a uniform confidence score of 1 can provide a more effective training signal for the network. In contrast, assigning varying confidence scores based on the distance from the apex may potentially introduce bias during the network learning process. Additionally, we believe that there is still a large scope for exploration and further research in pseudo-labeling methods.

Network input streams: The intuition behind the multi-stream network inputs lies in the multiple components of scene flow features. To explore this further, an experiment is conducted to compare the performance of using a multi-stream with individual components as input versus using a single-stream that incorporates all three components as input. As shown in Table 8, the results indicate that the performance achieved by utilizing the multi-stream is superior to using a single-stream across all tasks. This can be attributed to the different components of scene flow, which capture various aspects of motion, including horizontal, vertical, and depth information. By incorporating each component separately, the network can extract and leverage more fine-grained information about the ME movements, leading to improved performance.

Table 8: Performance comparison of using different network input streams with scene flow as input. Multi: multi-stream, Single: single-stream

Network input		Multi (Ours)	Single
Spotting (F1-score)	ME	**0.0695**	0.0426
	MaE	**0.0854**	0.0785
	Overall	**0.0809**	0.0656
ME Recognition (UF1)	4-class	**0.4006**	0.3575
	7-class	**0.2000**	0.1815
ME Analysis (STRS)	4-class	**0.0289**	0.0160
	7-class	**0.0164**	0.0082

6. Conclusion

Traditionally, facial MEs have been studied using 2D video sequences without facial depth information. Lately, the release of multi-modal datasets has made the development of 3D-based approaches possible. As one of the earliest works attempts on the multi-modal CAS(ME)^3 dataset, we propose an attention-based end-to-end multi-stream multi-task network with scene flow features as inputs, namely SFAMNet, which is demonstrated on three tasks: ME spotting; ME recognition; ME analysis. Specifically, the RGB-D flow is employed to compute the scene flow, which is an extension of optical flow with additional depth dimension. The experiment results prove that the scene flow is robust in estimating the 3D motion features, thus improving the network performance. Furthermore, we present a data augmentation strategy with a sliding window approach to enlarge the ME sample size for network training. Intuitively, this technique also stables the network performance towards the spotting and recognition tasks. To the best of our knowledge, this is the first known work that proposes an end-to-end training process to perform a one-stage ME analysis. We validate our proposed approach on the CAS(ME)^3 dataset, outperforming the benchmark results in all tasks. Further ablation studies are provided to discuss the interesting findings.

For future works, we plan to improve the RGB-D flow algorithm by adding occlusion handling to deal with "in-the-wild" scenarios, enhancing the ability to overcome the missing objects when a large motion such as MaE and head movement is present. Besides, the imbalanced dataset should be investigated to ensure the emotion labels in training data are distributed equally. While the multi-modal ME samples are still considered limited, ME generation techniques can be performed to increase the samples of the minority class. Additionally, the unlabeled data (part B in CAS(ME)^3 dataset) can be utilized for self-supervised learning. This learning paradigm forces the network to learn more meaningful visual clues before training on the actual tasks. It is worth studying the third-generation ME data (part C in CAS(ME)^3 dataset) with high ecological validity. This could benefit lie detection in crime scenes and offer a better understanding of ME occurrences.

Acknowledgment

This work was funded by Ministry of Science and Technology (MOST) (Grant Number: MOST 111-2221-E-035-059-MY3), the Ministry of Higher Education (MOHE) Malaysia under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2022/ICT02/HWUM/02/1), and HWUM JWS 2022/01.

References

[1] X. Liu, L. Jin, X. Han, J. You, Mutual information regularized identity-aware facial expression recognition in compressed video, Pattern Recognition 119 (2021) 108105.

[2] K. Yu, Z. Wang, L. Zhuo, J. Wang, Z. Chi, D. Feng, Learning realistic facial expressions from web images, Pattern Recognition 46 (8) (2013) 2144–2155.

[3] P. Ekman, Darwin, deception, and facial expression, Annals of the new York Academy of sciences 1000 (1) (2003) 205–221.

[4] P. Ekman, Microexpression training tool (METT). university of california, san francisco (2002).

[5] M. Frank, M. Herbasz, K. Sinuk, A. Keller, C. Nolan, I see how you feel: Training laypeople and professionals to recognize fleeting emotions, in: The Annual Meeting of the International Communication Association. Sheraton New York, New York City, 2009, pp. 1–35.

[6] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, V. M. Baskaran, A survey of automatic facial micro-expression analysis: databases, methods, and challenges, Frontiers in psychology 9 (2018) 1128.

[7] L. Yin, X. Wei, Y. Sun, J. Wang, M. J. Rosato, A 3d facial expression database for facial behavior research, in: 7th international conference on automatic face and gesture recognition (FGR06), IEEE, 2006, pp. 211–216.

[8] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, X. Fu, CAS(ME)^3: A Third Generation Facial Spontaneous Micro-Expression Database with Depth Information and High Ecological Validity, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[9] M. H. Yap, J. See, X. Hong, S.-J. Wang, Facial micro-expressions grand challenge 2018 summary, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 675–678.

[10] J. See, M. H. Yap, J. Li, X. Hong, S.-J. Wang, MEGC 2019–the second facial micro-expressions grand challenge, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[11] J. Li, S.-J. Wang, M. H. Yap, J. See, X. Hong, X. Li, MEGC 2020-the third facial micro-expression grand challenge, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 777–780.

[12] J. Li, M. H. Yap, W.-H. Cheng, J. See, X. Hong, X. Li, S.-J. Wang, FME 21: 1st Workshop on Facial Micro-Expression: Advanced Techniques for Facial Expressions Generation and Spotting, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5700–5701.

[13] J. Li, M. H. Yap, W.-H. Cheng, J. See, X. Hong, X. Li, S.-J. Wang, A. K. Davison, Y. Li, Z. Dong, MEGC 2022: ACM Multimedia 2022 Micro-Expression Grand Challenge, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7170–7174.

[14] G.-B. Liong, J. See, C.-S. Chan, Spot-then-recognize: A micro-expression analysis network for seamless evaluation of long videos, Signal Processing: Image Communication (2022) 116875.

[15] C. A. Corneanu, M. O. Simón, J. F. Cohn, S. E. Guerrero, Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications, IEEE transactions on pattern analysis and machine intelligence 38 (8) (2016) 1548–1568.

[16] X. Li, S. Cheng, Y. Li, M. Behzad, J. Shen, S. Zafeiriou, M. Pantic, G. Zhao, 4DME: A spontaneous 4d micro-expression dataset with multimodalities, IEEE Transactions on Affective Computing (2022).

[17] P. Ekman, W. V. Friesen, Nonverbal leakage and clues to deception, Psychiatry 32 (1) (1969) 88–106.

[18] S. Polikovsky, Y. Kameda, Y. Ohta, Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor (2009).

[19] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg), IEEE, 2013, pp. 1–6.

[20] A. Moilanen, G. Zhao, M. Pietikäinen, Spotting rapid facial movements from videos using appearance-based feature difference analysis, in: 2014 22nd international conference on pattern recognition, IEEE, 2014, pp. 1722–1727.

[21] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, CASME II: An improved spontaneous micro-expression database and the baseline evaluation, PloS one 9 (1) (2014) e86041.

[22] A. K. Davison, C. Lansley, N. Costen, K. Tan, M. H. Yap, SAMM: A spontaneous micro-facial movement dataset, IEEE transactions on affective computing 9 (1) (2016) 116–129.

[23] Y. Wang, J. See, R. C.-W. Phan, Y.-H. Oh, LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition, in: Asian conference on computer vision, Springer, 2014, pp. 525–537.

[24] X. Huang, G. Zhao, Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern, in: 2017 international conference on the frontiers and advances in data science (FADS), IEEE, 2017, pp. 159–164.

[25] B. K. Horn, B. G. Schunck, Determining optical flow, Artificial intelligence 17 (1-3) (1981) 185–203.

[26] B. D. Lucas, T. Kanade, et al., An iterative image registration technique with an application to stereo vision, Vol. 81, Vancouver, 1981.

[27] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Scandinavian conference on Image analysis,

Springer, 2003, pp. 363–370.

[28] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime TV-L1 optical flow, in: Joint pattern recognition symposium, Springer, 2007, pp. 214–223.

[29] M. Shreve, J. Brizzi, S. Fefilatyev, T. Luguev, D. Goldgof, S. Sarkar, Automatic expression spotting in videos, Image and Vision Computing 32 (8) (2014) 476–486.

[30] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, X. Fu, A main directional mean optical flow feature for spontaneous micro-expression recognition, IEEE Transactions on Affective Computing 7 (4) (2015) 299–310.

[31] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, M. Pietikäinen, Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods, IEEE transactions on affective computing 9 (4) (2017) 563–577.

[32] S.-T. Liong, J. See, K. Wong, R. C.-W. Phan, Less is more: Micro-expression recognition from video using apex frame, Signal Processing: Image Communication 62 (2018) 82–92.

[33] Y. He, S.-J. Wang, J. Li, M. H. Yap, Spotting macro-and micro-expression intervals in long video sequences, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 742–748.

[34] J. Li, C. Soladie, R. Seguier, LTP-ML: Micro-expression detection by recognition of local temporal pattern of facial movements, in: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 634–641.

[35] G.-B. Liong, J. See, L.-K. Wong, Shallow optical flow three-stream CNN for macro-and micro-expression spotting from long videos, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 2643–2647.

[36] J. Li, C. Soladie, R. Seguier, Local temporal pattern and data augmentation for micro-expression spotting, IEEE Transactions on Affective Computing (2020).

[37] Y. He, Research on micro-expression spotting method based on optical flow features, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4803–4807.

[38] Y. Zhao, X. Tong, Z. Zhu, J. Sheng, L. Dai, L. Xu, X. Xia, Y. Jiang, J. Li, Rethinking optical flow methods for micro-expression spotting, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7175–7179.

[39] J. Yu, Z. Cai, Z. Liu, G. Xie, P. He, Facial expression spotting based on optical flow features, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7205–7209.

[40] M. Verburg, V. Menkovski, Micro-expression detection in long videos using optical flow and recurrent neural networks, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–6.

[41] S.-J. Wang, Y. He, J. Li, X. Fu, Mesnet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos, IEEE Transactions on Image Processing 30 (2021) 3956–3969.

[42] W.-W. Yu, J. Jiang, Y.-J. Li, Lssnet: A two-stream convolutional neural network for spotting macro-and micro-expression in long videos, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4745–4749.

[43] W. Leng, S. Zhao, Y. Zhang, S. Liu, X. Mao, H. Wang, T. Xu, E. Chen, Abpn: Apex and boundary perception network for micro-and macro-expression spotting, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7160–7164.

[44] S. Wang, J. Yang, Z. Gao, Q. Ji, Feature and label relation modeling for multiple-facial action unit classification and intensity estimation, Pattern Recognition 65 (2017) 71–81.

[45] Y. Liu, H. Du, L. Zheng, T. Gedeon, A neural micro-expression recognizer, in: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), IEEE, 2019, pp. 1–4.

[46] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, Y.-C. Huang, Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[47] L. Zhou, Q. Mao, L. Xue, Dual-inception network for cross-database micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[48] L. Zhou, Q. Mao, X. Huang, F. Zhang, Z. Zhang, Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition, Pattern Recognition 122 (2022) 108275.

[49] J. Yu, C. Zhang, Y. Song, W. Cai, ICE-GAN: identity-aware and capsule-enhanced GAN for micro-expression recognition and synthesis (2020).

[50] S.-T. Liong, J. See, K. Wong, R. C.-W. Phan, Automatic micro-expression recognition from long video using a single spotted apex, in: Asian conference on computer vision, Springer, 2016, pp. 345–360.

[51] E. Herbst, X. Ren, D. Fox, RGB-D flow: Dense 3-d motion estimation using color and depth, in: 2013 IEEE international conference on robotics and automation, IEEE, 2013, pp. 2276–2282.

[52] Z. Yan, X. Xiang, Scene flow estimation: A survey, arXiv preprint arXiv:1612.02590 (2016).

[53] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, Y.-J. Liu, Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[54] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, X. Fu, How fast are the leaked facial expressions: The duration of micro-expressions, Journal of Nonverbal Behavior 37 (4) (2013) 217–230.

[55] B. Yang, J. Wu, Z. Zhou, M. Komiya, K. Kishimoto, J. Xu, K. Nonaka, T. Horiuchi, S. Komorita, G. Hattori, et al., Facial action unit-based deep learning framework for spotting macro-and micro-expressions in long video sequences, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4794–4798.

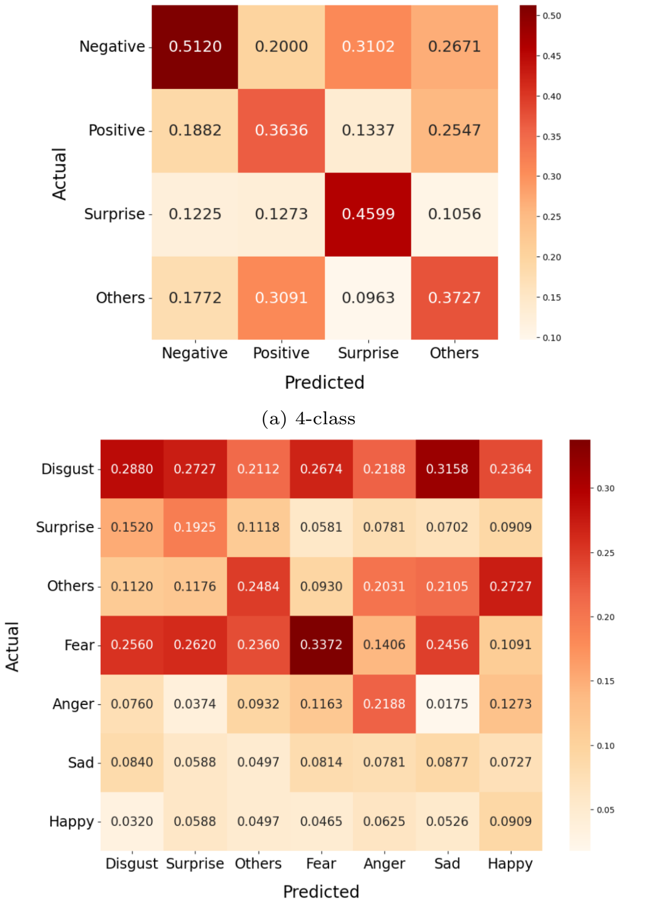
[56] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[57] G. B. Liong, S.-T. Liong, J. See, C.-S. Chan, MTSN: A Multi-Temporal Stream Network for Spotting Facial Macro-and Micro-Expression with Hard and Soft Pseudo-labels, in: Proceedings of the 2nd Workshop on Facial Micro-Expression: Advanced Techniques for Multi-Modal Facial Expression Analysis, 2022, pp. 3–10.

[58] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE signal processing letters 23 (10) (2016) 1499–1503.

[59] L.-W. Zhang, J. Li, S.-J. Wang, X.-H. Duan, W.-J. Yan, H.-Y. Xie, S.-C. Huang, Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences, in: 15th IEEE FG, 2020, pp. 245–252.

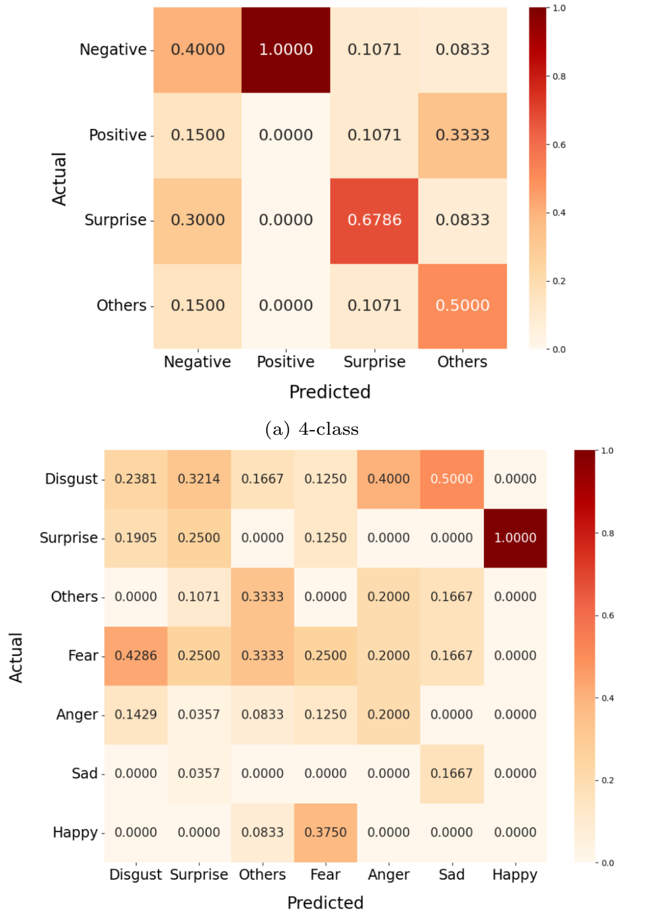
[60] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, G. Zhao, Revealing the invisible with model and data shrinking for composite-database micro-expression recognition, IEEE Transactions on Image Processing 29 (2020) 8590–8605.



(a) 4-class

(b) 7-class

Figure 5: Confusion matrices of ME recognition on CAS(ME)^3 dataset.



(a) 4-class

(b) 7-class

Figure 6: Confusion matrices of ME analysis on CAS(ME)^3 dataset.

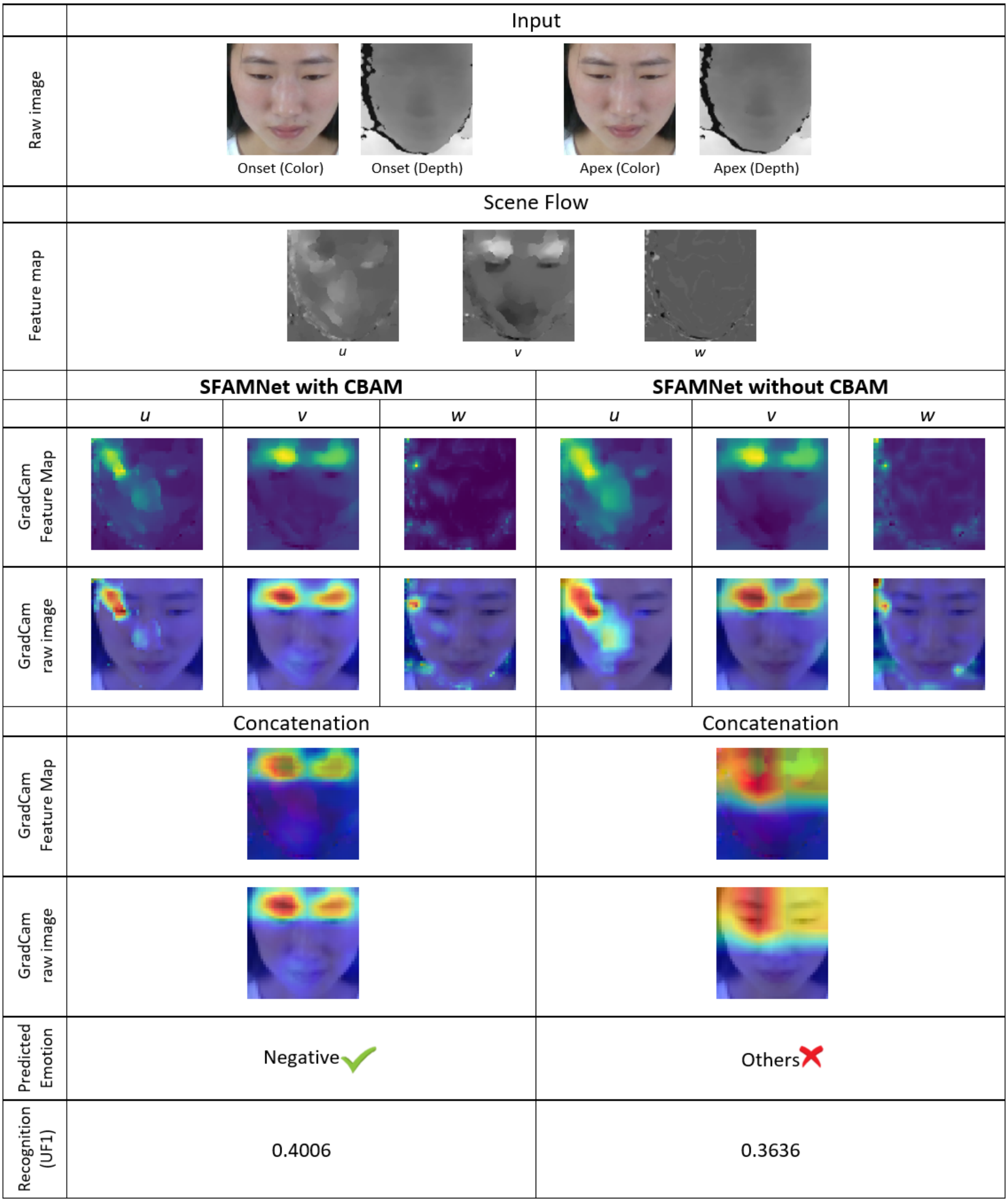


Figure 7: Gradcam visualization of the SFAMNet with and without the integration of CBAM. Taking video `sp153_a` in CAS(ME)^3 as an example, the activated regions from the network with CBAM are more precise, thus obtaining a higher recognition result on 4-class evaluation.